

Exercise – Data Processing using AWS Glue

Step 1: Create a new bucket so that AWS Glue can read from and write to the bucket in to it's own folders.

Create bucket

Buckets are containers for data stored in S3. [Learn more](#)

General configuration

Bucket name

myawsglueinputpouputbucket

Bucket name must be unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#)

AWS Region

Asia Pacific (Mumbai) ap-south-1

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

Choose bucket

Step 2: Create a couple of folders one for reading and one for writing to/from Glue. Upload the movielens csv file to the read directory in S3.

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) access your objects, you'll need to explicitly grant them permissions. [Learn more](#)



Copy S3 URI



Copy URL



Download

Create folder



Upload



Find objects by prefix



Name



Type



Last modified



read/

Folder

-



write/

Folder

-

Step 3: Go to AWS Glue in Amazon console

AWS Glue

Data catalog

Databases

Tables

Connections


Crawlers

Classifiers

Schema registries

Tables A table is the metadata definition that represents your data, including its location and schema.

[Add tables](#) [Action](#)

<input type="checkbox"/>	Name	Database
 You don't have any tables yet.		

Step 4: Click on Crawlers on the left panel and add a new Crawler

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries


Schemas

Settings

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

[Add crawler](#) [Run crawler](#) [Action](#) Showing: 0 - 0

<input type="checkbox"/>	Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
 You don't have any crawlers yet.								

Step 5: Give your crawler a name like S3-Crawler. Click Next

- ☒ Crawler info
- ☐ Crawler source type
- ☐ Data store
- ☐ IAM Role
- ☐ Schedule
- ☐ Output
- ☐ Review all steps

Add information about your crawler

Crawler name

► Tags, description, security configuration, and classifiers (optional)

[Next](#)

Step 6: Use the default options and click Next

- ✓ Crawler info
 - S3-Crawler
- Crawler source type
- Data store
- IAM Role
- Schedule
- Output
- Review all steps

Specify crawler source type

Choose Existing catalog tables to specify catalog tables as the crawler source. The selected tables specify the data stores to crawl. This option doesn't support JDBC data stores.

Crawler source type

☒ Data stores

☐ Existing catalog tables

Repeat crawls of S3 data stores

☒ Crawl all folders

Crawl all folders again with every subsequent crawl.

☐ Crawl new folders only

Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.

[Back](#) [Next](#)

Step 7: Add a data store by selecting the read directory in S3. Click Next

- ✓ Crawler info
 - S3-Crawler
- ✓ Crawler source type
 - Data stores
- Data store
- IAM Role
- Schedule
- Output
- Review all steps

Add a data store

Choose a data store

S3

Connection

Select a connection

Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any future S3 targets will also use the same connection (or none, if left blank).

[Add connection](#)

Crawl data in

☒ Specified path in my account

☐ Specified path in another account

Include path

s3://myawsglueinputpouputbucket/read

All folders and files contained in the include path are crawled. For example, type

Step 8: Select no in “Add another data store” and click Next

- ✓ Crawler info
 - S3-Crawler
- ✓ Crawler source type
 - Data stores
- Data store
 - S3: s3://myawsglue...
- IAM Role
- Schedule
- Output
- Review all steps

Add another data store

☐ Yes

☒ No

[Back](#) [Next](#)

Step 9: Create a new role in IAM and click Next

The screenshot shows the 'Choose an IAM role' step in the AWS Glue console. On the left, a sidebar lists the setup steps: Crawler info, Crawler source type, Data store, IAM Role (selected), Schedule, Output, and Review all steps. The main content area is titled 'Choose an IAM role' and explains that the IAM role allows the crawler to run and access Amazon S3 data stores. It provides three options: 'Update a policy in an IAM role', 'Choose an existing IAM role', and 'Create an IAM role' (which is selected). Below these options, the 'IAM role' section shows a text input field with 'glue-role' entered. A note states that to create an IAM role, the user must have 'CreateRole', 'CreatePolicy', and 'AttachRolePolicy' permissions. It then instructs the user to create an IAM role named 'AWSGlueServiceRole-rolename' and attach the AWS managed policy 'AWSGlueServiceRole' and an inline policy for read access to 's3://myawsglueinputpouputbucket/read'. At the bottom, there are 'Back' and 'Next' buttons.

Choose an IAM role

The IAM role allows the crawler to run and access your Amazon S3 data stores. [Learn more](#)

☐ Update a policy in an IAM role

☐ Choose an existing IAM role

☒ Create an IAM role

IAM role ⓘ

AWSGlueServiceRole- glue-role

To create an IAM role, you must have **CreateRole**, **CreatePolicy**, and **AttachRolePolicy** permissions.

Create an IAM role named "AWSGlueServiceRole-rolename" and attach the AWS managed policy, **AWSGlueServiceRole**, plus an inline policy that allows read access to:

- s3://myawsglueinputpouputbucket/read

You can also create an IAM role on the [IAM console](#).

[Back](#) [Next](#)

Step 10: Select the frequency as “Run On Demand” and click Next

The screenshot shows the 'Create a schedule for this crawler' step in the AWS Glue console. The left sidebar is updated to show 'IAM Role' as completed and 'Schedule' as the current step. The main content area is titled 'Create a schedule for this crawler' and features a 'Frequency' dropdown menu with 'Run on demand' selected. At the bottom, there are 'Back' and 'Next' buttons.

Create a schedule for this crawler

Frequency

Run on demand

[Back](#) [Next](#)

Step 11: Click Add database and enter a database of your choice, click create . Click Next

Add database

Database name

► **Description and location (optional)**

Resource link name

Shared database suggestions

Shared database

Shared database owner account ID

Create

Step 12: Review the steps and click Finish

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries

Schemas

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

[User preferences](#)

Add crawler **Run crawler** Action Filter by tags and attributes Showing: 1 - 1

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
S3-Crawler		Ready		0 secs	0 secs	0	0

Step 13: Select the crawler created and click Run crawler and wait for it to complete till you see status as ready and Tables added as 1

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata table in your data catalog.

[User preference](#)

Add crawler **Run crawler** Action Filter by tags and attributes Showing: 1 - 1

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
S3-Crawler		Ready	Logs	56 secs	56 secs	0	1

Step 14: Click database on the left panel and check if the database is created. Also click table to check if your table is created.

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries

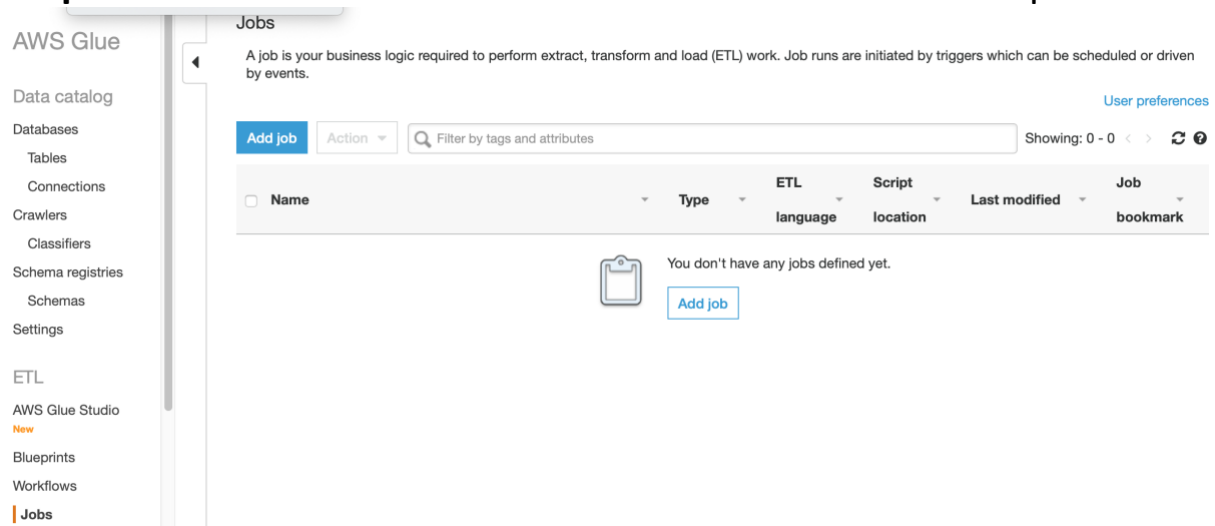
Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source

Add tables Action Filter by attributes or search by keyword **Save view**

Name	Database	Location	Classification	Last
read	s3-database	s3://myawsglueinputpo...	csv	6 J

Step 15: Click on the Jobs under the ETL section on the left panel



Step 16: Click on Add job and enter the details

Name : MyS3-ETL-Job

IAM Role: Select the available role

Under Security configuration modify

No of workers =2

Job timeout =10

Click Next

Step 17: Select your data source. Click Next

- ✔ Job properties
MyS3-ETL-Job
- Data source
- Transform type
- Data target
- Schema

Choose a data source

Filter by attributes or search by keyword

Showing: 1 - 1 < >

Name	Database	Location	Classification
read	s3-database	s3://myawsglueinputpouputbuck...	csv

Step 18: Select your Transformation type. Choose default. Click Next

- ✔ Job properties
MyS3-ETL-Job
- ✔ Data source
read
- Transform type
- Data target
- Schema

Choose a transform type

Machine learning transforms are currently not supported for Glue 2.0.

☒ Change schema
Change schema of your source data and create a new target dataset

☐ Find matching records
Use machine learning to find matching records within your source data

[Back](#) [Next](#)

Step 19: Select your data target. Choose the available option .Click Next

- ✔ Job properties
MyS3-ETL-Job
- ✔ Data source
read
- ✔ Transform type
Change schema
- Data target
- Schema

Choose a data target

☐ Create tables in your data target

☒ Use tables in the data catalog and update your data target

Filter by attributes or search by keyword

Showing: 1 - 1

Name	Database	Location	Classification
read	s3-database	s3://myawsglueinputpouputbucket/...	csv

Step 20: Select your data target. Choose the available option. Click Next

✓ Job properties
MyS3-ETL-Job

✓ Data source
read

✓ Transform type
Change schema

○ Data target
read

○ Schema

Choose a data target

☐ Create tables in your data target
☒ Use tables in the data catalog and update your data target

Filter by attributes or search by keyword

Showing: 1 - 1 < >

Name	Database	Location	Classification
read	s3-database	s3://myawsglueinputpouputbucket/...	csv

Step 21: Check the output schema definition and click Save job and edit script

✓ Job properties
MyS3-ETL-Job

✓ Data source
read

✓ Transform type
Change schema

✓ Data target
read

○ Schema

Output Schema Definition

Verify the mappings created by AWS Glue. Change mappings by choosing other columns with **Map to target**. You can **Clear** all mappings and **Reset** to default AWS Glue mappings. AWS Glue generates your script with the defined mappings.

Source	Target
Column name	Data type
rank	bigint
movie_title	string
year	bigint
rating	double

Map to target	Column name	Data type
rank	rank	long
movie_title	movie_title	string
year	year	long
rating	rating	double

Add column Clear Reset

Step 22: Replace the script with the one in Code/Glue/
glue_transformation.py and Click Save. Gibe the correct database,
input folder and output folder names in the script. Click Run Job.

Step 22.1: If you see an error “Job Failed”, the most probable reason would be that Glue does not have access to S3 to write data. So, go to IAM and in roles, search for the role you created in the Step 9 and add a policy S3Full Access and run the Job again.

Step 23: Check the job status and make sure it is success.

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries

Schemas

Settings

ETL

AWS Glue Studio

New

Blueprints

Workflows

my job

User preferences

Add job Action Filter by tags and attributes Showing: 1 - 1

Name	Type	ETL language	Script location	Last modified	Job bookmark
my job	Spark	python	s3://aws-gl...	6 June 2021 5:34 P...	Disable

Run ID	Retry attempt	Run status	Error	Output Logs	Error logs	Glue version	Maxin Triggered by	Start time	End time	Start-up time	Execu time	Timeec Delay	Job run input
jr_93f6fa7...	-	Succ eede d		Logs	Error logs	2.0	2	6 J...	6 J...	7 secs	1 min	10 mins	s3://aws-...

Step 24: Go to your bucket and check the output directory. It should contain a file which has the results

Adding a Workflow

Step 1: On the left panel click Work flow. Click Add workflow

The screenshot shows the AWS Glue Studio interface. On the left sidebar, the 'Workflows' option is highlighted under the 'ETL' section. The main panel displays the 'Workflows (0)' page. At the top, there is a description: 'A workflow is an orchestration used to visualize and manage the relationship and execution of multiple triggers, jobs and crawlers.' Below this, there is a blue 'Add workflow' button, an 'Actions' dropdown menu, and a refresh icon. A search bar with the placeholder 'Filter workflows' is also present. Below the search bar is a table with columns: 'Name', 'Last run', 'Last run status', and 'Last modified'. The table is currently empty, and a message 'No workflows' is displayed. Below the message is a blue button labeled 'Add a new ETL workflow'.

Step 2: Enter work flow name and click Add Workflow

The screenshot shows the 'Add a new ETL workflow' form in AWS Glue Studio. The left sidebar shows the 'Workflows' option highlighted. The main panel has a breadcrumb 'Workflows > Add workflow' and the title 'Add a new ETL workflow'. Below the title is a description: 'Add a workflow in order to orchestrate ETL jobs, triggers, and crawlers'. The form has three sections: 1. 'Workflow name' with a text input field containing 'MyWorlFlow' and a note: 'Names may only contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_)'. 2. 'Description (optional)' with a text area containing the placeholder 'Enter a workflow description...' and a note: '250 characters max'. 3. 'Default run properties (optional)' with the text 'No default run properties' and a blue 'Add property' button.

Step 3: Click Add Trigger

✓ Successfully created workflow: MyWorlFlow.

Workflows (1)

A workflow is an orchestration used to visualize and manage the relationship and execution of multiple triggers, jobs and crawlers.

Add workflow

Actions

Filter workflows

< 1 >

Name	Last run	Last run status	Last modified
MyWorlFlow	-	-	Mon, 27 Dec 2021 17:29:38 G...

Graph

Details

History

Legend:

● Start

◆ Trigger

▣ Job

▣ Crawler

🔴 Incomplete

🔴 Error

🗑 Deleting

Remove

Action

The workflow is empty

Add trigger

Step 4: Give trigger a name and select OnDemand for trigger type and click Add

Add trigger

Clone existing

Add new

Name

StartCrawler

Description (optional)

Type description...

Trigger type

Cancel

Add

Step 5: Select the Node which pops up a window to add a crawler. Select the crawler tab and select the crawler. Click Add

Add job(s) and crawler(s) to trigger

Jobs

Crawlers

Filter crawlers

< 1 >

<input checked="" type="checkbox"/>	Name	Description
<input checked="" type="checkbox"/>	S3-crawler	

Cancel

Add

Step 6: Click Actions and select Add Trigger. Give a job name and click Add

Add trigger

Clone existing

Add new

Name

StartJob

Description (optional)

Type description...

Trigger type

☐ Schedule ☒ Event ☐ On demand ☐ EventBridge event

Trigger logic

☒ Start after ANY watched event ☐ Start after ALL watched event

Cancel

Add

Step 7: Select the Any trigger and click actions and select Add Jobs/crawlers to watch. Select your crawler under the Crawlers and click Add

Add job(s) and crawler(s) to watch ×

Jobs

Crawlers

< 1 >

<input checked="" type="checkbox"/>	Name	Description
<input checked="" type="checkbox"/>	S3-crawler	

Crawler event to watch

SUCCEEDED

Cancel

Add

Step 8: Click on the Node after the Start Job and select your job under the Jobs Tab. Click Add

Add job(s) and crawler(s) to trigger ×

Jobs

Crawlers

< 1 >

<input checked="" type="checkbox"/>	Name	Type	Last modified
<input checked="" type="checkbox"/>	MyJob	Spark	Mon Dec 27 2021 22:24:34 GMT+0530 (IST)

Cancel

Add

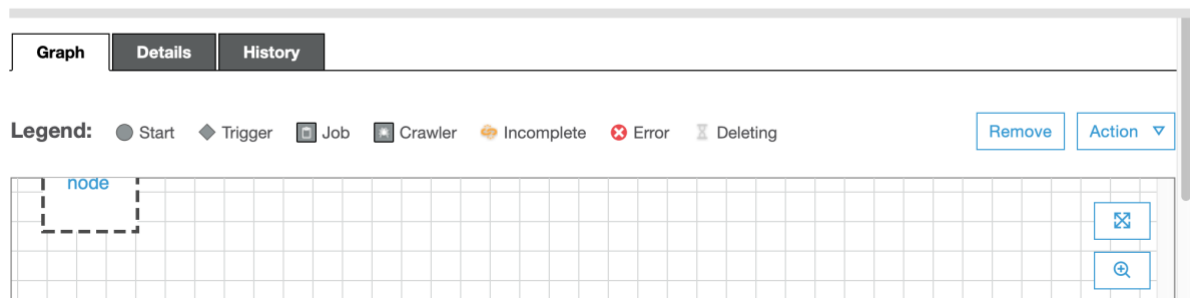
Step 9: From the actions menu, click on Run

✔ Successfully created workflow: MyWorlFlow. ✕

Workflows (1)

A workflow is an orchestration used to visualize and manage the relationship and execution of multiple triggers, jobs and crawlers.

Add workflow		Actions ▾		<input type="text" value="Filter workflows"/>	< 1 >			
	Name	▼	Last run	▼	Last run status	▼	Last modified	▼
	MyWo	<div>Delete Run Edit</div>	-		-		Mon, 27 Dec 2021 17:29:38 G...	



Step 10: Wait till the works goes to completed state and click the output directory in S3

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Schema registries

Schemas

Settings


✔ Successfully started workflow: MyWorlFlow. ✕

Workflows (1)

A workflow is an orchestration used to visualize and manage the relationship and execution of multiple triggers, jobs and crawlers.

Add workflow


Actions



🔍

Filter workflows

< 1 >

	Name	Last run	Last run status	Last modified
	MyWorlFlow	Mon, 27 Dec 2021 18:...	Completed	Mon, 27 Dec 2021 17:29:38 G...